Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator

Stephen A. Klein[1], Yuying Zhang[1], Mark D. Zelinka[1], Robert Pincus[2], James Boyle[1], and Peter J. Gleckler[1]

[1]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California, USA

[2]University of Colorado and NOAA/Earth System Research Laboratory, Boulder, Colorado, USA

Submitted to *Journal of Geophysical Research – Atmospheres*, July 2012; Revised, xxxx 2012; Accepted, xxxx 2012

Corresponding author: S. A. Klein, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, 7000 East Avenue, L-103, Livermore, CA 94551. (klein21@llnl.gov)

Running Title: Evaluating Clouds in Climate Models

Key Points

- Newer climate models have improved simulations of cloud optical depth

- Cloud amount and cloud-top pressure simulations show smaller improvement
- Newer models have fewer compensating errors in their radiation budget

Abstract

The annual cycle climatology of cloud amount, cloud-top pressure and optical thickness in two climate model ensembles is compared to satellite observations in order to identify changes over time in the fidelity of climate model simulations of clouds. In more recent models, there is widespread reduction of a bias associated with too many highly reflective clouds, with the best models having eliminated this bias. With increased amounts of clouds with lesser reflectivity, recent models have reduced the compensating errors that permit models to simulate the time-mean radiation balance. Errors in cloud amount as a function of height or climate regime on average show little change or small improvement, although greater improvement can be found in the models of individual modeling centers.

Index Terms: 3337 Atmospheric Processes: Global climate models (1626, 4928); 3310 Atmospheric Processes: Clouds and cloud feedbacks; 3360 Atmospheric Processes: Remote sensing (4337)

Keywords: clouds, climate models, satellite simulator

1. Measuring changes in the simulations of global cloudiness over time

The representation of clouds by climate models is a key ongoing challenge in the numerical representation of Earth's climate. Due to their large impact on Earth's radiation budget, clouds are important for determining aspects of current climate, such as surface air temperatures in many regions [*Ma et al.*, 1996; *Curry et al.*, 1996], the strength and variability of atmospheric circulations [*Slingo and Slingo*, 1988], and the magnitude of climate changes that result from perturbations in the chemical composition of the atmosphere [*IPCC*, 2007]. While important, the modeling of clouds is very difficult because most cloud processes happen at scales far smaller than can be resolved by climate models, and thus their bulk effects must be represented with imperfect parameterizations.

A large effort of many scientists over several decades and on several fronts has been undertaken to improve our understanding of cloud processes, often with the ultimate goal of improving the modeling of clouds in climate models. Observational programs have been launched to better understand cloud processes [*Stephens et al.*, 2002; *Ackerman and Stokes*, 2003], while very high-resolution models capable of resolving cloud processes provide additional information for the development of cloud parameterizations unavailable from observations [*GEWEX Cloud System Science Team*, 1993]. The community of scientists that work on physical process parameterizations in climate models has used the information provided by observations and fine-scale models to develop and implement many new and improved cloud parameterizations. Cloud simulations in climate models may also be improved indirectly by complementary model development efforts that improve the representation of other physical processes including atmospheric dynamics, as well as by increases in model resolution.

Given this effort, it is important to ask: are climate model simulations of clouds improving and, if so, by how much? Here, we analyze the ability of two generations of climate models to simulate the climatological distribution of clouds and judge fidelity by comparison to several decade's worth of satellite observations. Because of the significant differences between the ways clouds are observed and the ways they are represented in models, we use a "satellite simulator" to increase the chances that differences between the models and observations represent actual model deficiencies. We find that significant progress in the ability of models to simulate clouds has occurred over the last decade, particularly in reducing the over-prediction of highly reflective clouds [*Zhang et al.*, 2005].

2. Climate Models, Satellite Observations, ISCCP Simulator and Analysis Methods

2.1 Climate Models

The models we analyze are those that submitted output to the first two phases of the Cloud Feedback Model Intercomparison Project [*McAvaney and LeTreut*, 2003; *Bony et al.*, 2011]. Submissions to the first phase (CFMIP1) were completed by the end of 2005 and thus the nine models (Table 1) we analyze were all formulated prior to that time, with HadSM3 being perhaps the oldest of these models. Submissions to the second phase (CFMIP2) began in late 2011 and as of the time of this writing[1] we have output from eight models (Table 2). CFMIP2 is a subset of the much wider fifth Coupled Model Intercomparison Project (CMIP5) [*Taylor et al.*, 2012] associated with the fifth assessment report of the Intergovernmental Panel on Climate Change. Although less formal, there was also a close connection between CFMIP1 and the corresponding third Coupled Model Intercomparison Project (CMIP3) [*Meehl et al.*, 2007]. As some models that participated in CFMIP1 did not participate in CMIP3, we retain the more accurate label of CFMIP (instead of CMIP) when referring to the ensembles.

A direct evaluation of model changes is complicated by the fact that the CFMIP1 output used here is from the control climate integrations of slab-ocean models (i.e., atmospheric models coupled with a mixed-layer model of the upper ocean), while the CFMIP2 output is from simulations of the atmosphere model with sea surface temperatures and sea-ice distributions prescribed from observations from recent decades (i.e. Atmospheric Model Intercomparison Project (AMIP) simulations [*Gates et al.*, 1999]). This difference arises because the satellite simulator output we require is only available from the slab-ocean models of CFMIP1, while the slab-ocean model framework is not part of CFMIP2. Nonetheless, we believe that the difference in modeling framework has only a minor

[1] We intend to add other CFMIP2 models to our analysis should they become available during the review process. We think it is possible to add results from the GFDL AM3 and EC-Earth models.

impact, because the differences in surface boundary conditions between slab-ocean models and AMIP integrations (and hence the resulting distribution of clouds) are small, even for slab-ocean models constructed to mimic the climate of the pre-industrial era. We have tested this notion by comparing AMIP and slab-ocean model simulations for one model (CAM4), and find that differences in our results resulting from the different modeling frameworks to be much smaller than differences among CFMIP models.

2.2 Satellite Observations

We compare the clouds simulated by climate models to the cloud climatology of observations created by the International Satellite Cloud Climatology Project (ISCCP) [*Rossow and Schiffer*, 1991, 1999]. ISCCP provides estimates of the area coverage of clouds stratified by *ctp*, the apparent cloud-top pressure of the highest cloud in a column, and by τ, the column integrated optical thickness of clouds. These estimates are the results of retrieval algorithms applied to radiance observations from the visible and infrared window channels of geostationary and polar orbiting satellites. They are accumulated for 280 km x 280 km regions every 3 hours staring in July 1983 and we use data through June 2008. Area coverage estimates are summarized in a joint histogram with 6 bins in τ and 7 bins in *ctp*; bin boundaries are shown in Figure 7. We use custom-built daytime-only monthly averages that are described more fully in *Pincus et al.* [2012] and are available from http://climserv.ipsl.polytechnique.fr/.

As a point of comparison, we also use roughly analogous observations from the

MOderate Resolution Imaging Spectrometer (MODIS) instruments for the period March 2000 through April 2011 [*Pincus et al.*, 2012]. MODIS uses substantially different methods of estimating *ctp* than does ISCCP, so the amounts of clouds in each bin of the joint histogram of *ctp* and τ from MODIS are not comparable to those observed by ISCCP or the output of an ISCCP simulator applied to climate models. (MODIS observations may be compared to the output of a MODIS simulator [*Pincus et al.,* 2012], but that was not available at the time of CFMIP1.) On the other hand, MODIS retrievals of τ are roughly equivalent to those from ISCCP, so we compare MODIS observations, aggregated over bins of *ctp*, to both ISCCP observations and the output of ISCCP simulators.

2.3 ISCCP Simulator

A satellite simulator is a diagnostic code applied to model variables that reduces the influences of inconsistencies between the ways clouds are observed and the ways they are modeled [*Bodas-Salcedo et al.,* 2011]. By mimicking the observational process in a simplified way, the simulator attempts to compute what a satellite would retrieve if the real-word atmosphere had the clouds of the model. Simulators increase the chances that the comparison of satellite retrievals to model output after run through a simulator is an evaluation of the fidelity of a model's simulation rather than a reflection of observational limitations or artifacts. The use of a satellite simulator also puts model intercomparison on a firmer basis by minimizing the impacts of how clouds are defined in different parameterizations.

The ISCCP simulator is the oldest of the satellite simulators used to evaluate clouds in models and has been widely used by most major climate modeling centers since its creation over ten years ago [*Klein and Jakob*, 1999; *Webb et al.*, 2001]. The ISCCP simulator mimics the key aspect of the ISCCP retrieval algorithms that radiances in every cloudy satellite pixel are assumed to arise from a single homogenous layer of cloud with *ctp* determined from an infrared brightness temperature. In detail, the ISCCP simulator takes a model's vertical profile of grid-box mean clouds and creates a set of sub-grid scale columns which are completely clear or cloudy at each level and which are consistent with the model's cloud-overlap parameterization. (This step is bypassed for models that provide to the simulator a set of previously generated sub-grid scale columns.) From every sub-grid scale column, one determines the single value of *ctp* and column-integrated τ that would be consistent with the single-layer cloud retrieval that ISCCP applies to every cloudy satellite pixel. In this step, *ctp* is determined by applying a simplified radiative transfer model in each sub-grid scale column to determine an infrared brightness temperature, which is then converted to the temperature at cloud-top by using a cloud longwave emissivity derived from τ, as in the ISCCP retrieval algorithm. Once a cloud-top temperature has been determined, *ctp* is equated with the interpolated pressure that has the identical temperature according to the model's profile of temperature. The column-integrated value of τ is equated with the sum of model-reported τ from all model layers that are cloudy in a given sub-grid scale column. From these sub-grid scale values of *ctp* and τ, the grid-box mean joint histogram of *ctp* and τ is formed for every grid box and then subsequently averaged over time. To make the comparison with satellite retrievals of τ more fair, the ISCCP simulator is only applied to grid-boxes that are sunlit at a given model time.

The ISCCP simulator itself changed between CFMIP1, which used v3.5, and CFMIP2, which used v4.1, raising the possibility that differences in the diagnostics might be mistaken for changes in simulation quality. The most significant algorithmic difference between these two versions involves the determination of *ctp* for clouds under atmospheric temperature inversions, such as subtropical marine stratocumulus. In these situations, ISCCP often erroneously assigns *ctp* to a level far higher (100 – 300 hPa) in the atmosphere than it should be [*Garay et al.*, 2008]. In CFMIP1, *ctp* is assigned to the highest interpolated pressure (lowest altitude) with matching cloud-top temperature, but, since the simulator is intended to mimic the retrieval process (even when it is faulty), the simulator was changed so that *ctp* is assigned to the lowest interpolated pressure (highest altitude) with matching cloud-top temperature when a temperature inversion is present in the model. We have verified that the impact of this and other simulator differences have little impact on our results by comparing the output of these two versions of the ISCCP simulator when applied to identical integrations of two CFMIP2 models (CAM4 and HadGEM2) (not shown). Simulator changes primarily affect *ctp* with differences of up to 0.01 in the amounts of clouds annually averaged over the domain 60°N-60°S for *ctp* bins where *ctp* < 680 hPa, and somewhat larger differences of up to 0.04 for *ctp* bins where *ctp* > 680 hPa.

We only use models for which we are reasonably confident of a correct implementation of the ISCCP simulator. Our primary test is to verify that the sum of cloud cover over all bins of the joint histogram is consistent with the model diagnostic of total cloud cover ('clt') which a model computes without using the ISCCP simulator [*Zelinka et al.*, 2012].

2.4 Analysis Methods

Climatological joint histograms of *ctp* and τ are formed for every calendar month by averaging model and observational data on a common 2° latitude by 2.5° longitude grid from every available year. Most model climatologies are based upon either 20 or 30 simulated years whereas the observed climatologies are for 25 years for ISCCP and 11 years for MODIS, but differences in the number of years available do not materially affect our evaluation [*Pincus et al.*, 2008]. (The scalar measures of the fidelity of model simulations [Section 4] are sensitive to this issue if the number of years used to form a climatology is very low (< 5); this only affects results for the two MIROC models in CFMIP1.) To minimize issues with cloud retrievals above surfaces with snow or ice, we restrict our analysis to the domain 60°N-60°S.

We evaluate changes over time in two ways. One considers changes in the multi-model mean from each of the CFMIP ensembles. This has the advantage of considering all available models and of highlighting common model errors. However, multi-model means are sensitive to the addition of new models (especially given the small sizes of the model ensembles) and changes in the multi-model mean may not reveal individual model error reductions when the spread of model results is centered on the observed value, as is often the case [*Gleckler et al.*, 2008]. To address these limitations, we also track the changes over time in the models from the four modeling centers that have contributed one or more models to both ensembles. For this analysis, we use models from the Canadian Centre for Climate Modeling and Analysis (AGCM4.0 to CanAM4), the United Kingdom's Met Office Hadley Centre (HadSM3 to HadSM4 to HadGEM1 to HadGEM2), the Japanese climate model effort associated with MIROC (MIROC(hisens) and MIROC(losens) to MIROC5), and the United States climate modeling effort associated with the Community Atmosphere Model (CCSM3.0 to CAM4 to CAM5).

3. Comparisons of climate model simulations of clouds to satellite observations

3.1 Common improvements and failures in the simulation of total cloud amount

We begin our analysis by examining the ability of models to simulate the space-time distribution of total cloud amount, i.e., how often a cloud occurs with any value of *ctp* and τ, which is perhaps the most fundamental aspect of a model's ability to simulate clouds. Unfortunately, this quantity is problematic to define from observations: satellite estimates of total cloud amount are extremely sensitive to many observational factors including the scale and sensitivity of the fundamental observations, as well as decisions made during the aggregation to larger scales [*Stubenrauch et al.*, 2009; *Mace et* *al.*, 2009; *Marchand et al.*, 2010; *Pincus et al.,* 2012]. We make the comparison more robust by restricting the analysis to clouds with τ exceeding some minimum threshold τ_{min}, which we set to minimize hard-to-detect and partly-cloudy observations. We select τ_{min} = 1.3 from among the discrete choices offered by the bin boundaries of the joint histogram of *ctp* and τ by balancing the following desires: (a) to maximize the number of clouds that we examine, (b) so that the observational datasets we use agree among themselves, ensuring robust model evaluation, and (c) to minimize the chances that an observational platform would have missed a cloud with $\tau > \tau_{min}$. Setting τ_{min} = 1.3 provides the smallest relative bias and relative root-mean-square difference, as well as the maximum correlation coefficient, between the space-time distributions of the annual cycle climatologies of ISCCP and MODIS.

Figure 1 illustrates the annual mean total cloud amount for the multi-model means of the CFMIP1 and CFMIP2 ensembles, the ISCCP and MODIS observations, and the difference of the CFMIP2 multi-model mean with ISCCP observations and with the CFMIP1 multi-model mean. For the domain 60°N-60°S, the annual mean total cloud amount fraction with a τ_{min} of 1.3 from ISCCP and MODIS is 0.51 and 0.47, respectively. The multi-model means of both CFMIP1 and CFMIP2 are 0.43 with more than ¾ of models in both ensembles below the range of observational estimates. Although the multi-model mean is identical between the two ensembles, if one examines these area-averaged values for the four model families in which we can track progress, in every case the most recent model is closer to the observational estimates. The increase is quite striking for the Hadley Centre models, with HadSM3 having a total cloud amount of 0.33 but HadGEM2 having a total cloud amount of 0.43.

Relative to ISCCP observations, model underestimates of total cloud amount preferentially occur in regions of subtropical marine stratocumulus on the eastern sides of subtropical ocean basins and over middle latitudes. In the stratocumulus regions, there is a wide variety of results in both ensembles with about 3-4 members in each ensemble having total cloud amount values close to observed and the reminder of models significantly below observational estimates. Although the differences between the multi-model means of ensembles are small in these regions, one finds marked progress in 3 out of the 4 families we can track with the amount of clouds in the most recent model versions close to observed. This suggests that at least for the modeling centers for which we can track progress, the simulation of current climate amounts of subtropical stratocumulus has been improving, perhaps in response to the well-known importance of the low clouds in these regions for mean climate and climate sensitivity [*Bony and duFresne*, 2005].

Although not as well known, models also typically underestimate total cloud amount at middle latitudes over both land and ocean (Figure 1). While a few models are close to observed over the middle latitude oceans, all models underestimate total cloud amount over the middle latitudes of Eurasia and North America. Examination of level-by-level cloud amount indicates that these underestimates, over both land and ocean, are primarily of lower level clouds (*ctp* > 560 hPa) although underestimates in upper level clouds (*ctp* < 560 hPa) do contribute some to this error, depending on the model. When examining results by model families, one finds no consistent sign of progress for this bias either over ocean or land, consistent with larger middle-latitude bias in the CFMIP2 multi-model mean relative to the CFMIP1 multi-model mean.

3.2 Improvements as a function of cloud-top pressure and cloud optical depth

In addition to getting clouds to occur in the right places and times, having a good simulation of *ctp* and τ is essential to getting the correct long- and shortwave impacts of a given cloud on the top-of-atmosphere radiation budget. Figure 2 illustrates the amount of clouds with $\tau > 1.3$ as a function of *ctp* averaged over 60°N-60°S. Models tend to underestimate the amount of middle (440 hPa < *ctp* < 680 hPa) and low-level (*ctp* > 680 hPa) clouds while having about the right amount of high-level (*ctp* < 440 hPa) clouds [*Zhang et al.*, 2005]. The general underestimate of low-level clouds is consistent with the lack of clouds in marine stratocumulus and middle-latitudes mentioned above. Differences in middle-level clouds are somewhat hard to interpret as many middle-level clouds observed by ISCCP are in fact multi-layer cloud scenes of cirrus above boundary layer cloud [*Marchand et al.*, 2010; *Mace et al.*, 2011]. Though the ISCCP simulator is capable of reproducing this artifact [*Mace et al.*, 2011], it will do so only if a model produces thin cirrus over boundary layer clouds. Thus, underestimates of middle-level cloud may actually indicate a lack of cirrus above boundary layer cloud.

Relative to that of the CFMIP1 ensemble, the CFMIP2 multi-model mean is closer to the observed amounts for 6 out of 7 bins of *ctp*, suggesting some improvement. This improvement is noticeable in the relative amounts of low-level clouds in the two lowest *ctp* bins. While a large part of this improvement is due to the change in the simulator's determination of *ctp* for clouds under an inversion, improvement can be found in the models from modeling centers that contribute more than one model to a given ensemble (compare HadSM3 to HadGSM1 and CAM4 to CAM5). Because the ISCCP simulator version does not change within these two pairs, we can conclude that these models have improved their simulation of low-level clouds. For middle-level clouds, there is also a reduction in the model underestimate, particularly for the 560-680 hPa *ctp* bin. In fact, the perfect agreement of CAM5 with ISCCP for this bin can be attributed to the fact that snow is now radiatively active and thus the simulator counts the contribution of snow to τ and the infrared-brightness temperature used to determine *ctp* [*Kay et al.*, 2012].

Figure 3 illustrates the amount of clouds as a function of τ regardless of *ctp* and averaged over 60°N-60°S. More so than in the case of *ctp*, rather marked improvement can be seen for τ bins where ISCCP and MODIS agree fairly well ($\tau > 3.6$). In particular, the amounts of optically thick clouds ($\tau > 23$) are significantly closer to observed in the CFMIP2 ensemble relative to the CFMIP1 ensemble with a marked reduction in the previously identified overestimate of highly reflective clouds [*Zhang et al.*, 2005]. All but one of the CFMIP2 models have fewer clouds in the optically thickest bin ($\tau > 60$) than all but one of the CFMIP1 models. This bias reduction is widespread enough that it is dramatically present for each of the 4 model families in which we can track progress (Figure 4).

The fraction of the 60°N-60°S area covered by optically thick cloud is 0.175 for the CFMIP1 ensemble mean but is 0.121 for the CFMIP2 ensemble mean. The CFMIP2 ensemble mean is still larger than the observational estimates of 0.064 for ISCCP and 0.082 for MODIS, indicating that about half of the bias remains. For HadGEM2 and MRI-CGCM3, the amount of optically thick cloud is within the range of the two observational estimates. The reduction between ensembles in optically thick clouds is larger for lower-level (*ctp* > 560 hPa) clouds than it is for upper-level (*ctp* < 560 hPa) clouds, 0.043 vs. 0.009 respectively. With the greater reduction in lower-level optically thick clouds, 7 out of 8 CFMIP2 models as opposed to 5 out of 9 CFMIP1 models reproduce the fact that optically thick clouds occur more frequently with *ctp* at upper levels than at lower levels. However, for only 2 CFMIP1 and 3 CFMIP2 models does the ratio of upper to lower-level optically thick clouds exceed the observed value of 1.7 for ISCCP (2.2 for MODIS).

Geographically, one can see from the multi-model means that the significant reductions in the amount of optically thick clouds occur over both the subtropical stratocumulus regions and middle-latitude land and especially ocean (Figure 5). There is no improvement in the multi-model mean overestimate of optically thick clouds over tropical continents, and this bias is present in 7 out of 9 CFMIP1 models and 7 out of 8 CFMIP2 models. We suspect that the common model bias in the diurnal cycle precipitation over tropical land [*Yang and Slingo*, 2001; *Dai*, 2006] contributes to this error by producing too many optically thick anvil clouds near mid-day, when they are visible to the ISCCP simulator, rather than at night.

The decrease in optically thick clouds has been accompanied by an increase in the amount of clouds with intermediate optical depths ($3.6 < \tau < 23$) (Figures 3 and 6). This increase is present in each of the 4 model families for which we can track progress, with one 1 CFMIP model (IPSL-CM4) and 3 CFMIP2 models (CAM5, HadGEM2, MPI-ESM-LR) having the amount of intermediate optical depth clouds lying in between the values from ISCCP and MODIS.

Passive observational estimates of the amount of cloud with $0.3 < \tau < 3.6$ disagree sharply, in part because many of the observations which produce clouds in this optical thickness range are partly cloudy [*Pincus et al.,* 2012]. This makes it impossible to assess the fidelity of model simulations for these clouds. For $\tau < 0.3$, there is a wide variety of model results, particularly in CFMIP1 where the two MIROC models each have more than 0.25 of the area covered by clouds of this optical depth range. Clouds this thin have too little contrast on the top-of-atmosphere radiation budget to be detected with the passive sensors used by ISCCP and MODIS; in fact, the τ bin boundary of 0.3 is chosen to crudely mimic a sensitivity threshold for ISCCP (W. B. Rossow, personal communication). Assessment of very thin clouds requires the use of an active sensor such as CALIPSO [*Winker et al.*, 2009]. Such an assessment would be relevant for the plentiful but very thin tropopause-level cirrus in the tropics [*Mace et al.*, 2009; *Thorsen et* *al.*, 2011].

3.3 Radiative impact of model errors in cloud properties

As in nature, clouds in climate models strongly affect the radiation balance as a function of space and time. Model tuning guarantees that the global and annual average of the net radiation is close to zero, but significant regional errors in the radiation field may persist, and correct regional fluxes can be achieved through compensating errors in cloud properties. One common error is to have clouds which are too few but too bright, that is, to have lower-than-observed cloud amounts with larger-than-observed values of τ, such that the average shortwave radiation budget is about right [*Zhang et al.*, 2005].

We explore these issues by using cloud radiative kernels [*Zelinka et al.*, 2012] to compute the radiative effects of errors in cloud properties. A cloud kernel $K^{SW,LW}$ is the result of a radiative transfer calculation that computes the impact on the top-of-atmosphere short- and long-wave fluxes, relative to clear-sky, of the addition of a unit area covered by a cloud with a given *ctp* and τ. Our kernels are computed as a function of latitude, longitude and calendar month. Multiplying the kernels by the bias, relative to ISCCP, in cloud amount in each bin of the joint histogram yields an estimate of the error in top-of-atmosphere radiation budget due to errors in the simulated distribution of clouds as a function of *ctp* and τ.

Figure 7 shows the annually and 60°N- 60°S averaged bias relative to ISCCP in cloud amount fraction in the joint histograms of *ctp* and τ for the four model families in which we can track progress and the multi-model means for CFMIP1 and CFMIP2. Figure 8 and 9 show the corresponding biases in W m^{-2} for the short- and long-wave radiation of the same models. (The Canadian model pairing is absent from Figures 8-9 because we cannot perform accurate cloud kernel calculations for AGCM4.0 for the reasons discussed in the Appendix of *Zelinka et al.* [2012].) The oldest models are in the left column and the most recent models on the right. The prominent overestimate of optically thick clouds occurs in nearly all *ctp* bins in the earlier models, but is much reduced in the more recent set. Likewise the underestimate of optically thin ($0.3 < \tau < 3.6$) and intermediate clouds present in nearly all *ctp* bins has been reduced in the more recent model versions. As discussed above, whether or not the biases in thin clouds are real is unclear.

The radiative impact of these biases on the short-wave spectrum quantifies the nature of compensating errors (Figure 8), with the overestimates of reflected shortwave by clouds with $\tau > 23$ compensating for a lack of reflection by clouds with thin and intermediate optical depths. The figure is similar to that of the cloud biases (Figure 7) except that weighting by the shortwave radiative kernel reduces the impact of the underestimate of optically thin clouds relative to the overestimate of optically thick clouds. The degree of compensation is markedly reduced in the more recent models. For example, in HadSM3 there was 27 W m^{-2} too much reflectence by clouds with $\tau > 60$, whereas in the most recent model HadGEM2-A, the bias is less than 1 W m^{-2}. Similarly, for HadSM3, CCSM3.0 and MIROC (hisens) and MIROC (losens), there was an underestimate of reflected shortwave radiation by clouds with $3.6 < \tau < 9.4$ of about 10 W m^{-2}, but in CAM5, MIROC5, and HadGEM2-A this bias is less than 3 W m^{-2}. In the multi-model mean, too much reflectance by optically thick clouds compensates for an underestimate in reflection by clouds with $1.3 < \tau < 9.4$ for most *ctp* bins, but the biases are smaller in the more recent models.

In the longwave spectrum, the nature of compensating biases is similar but with emphasis on upper level clouds (Figure 9). In general, there is too much reduction of outgoing longwave radiation by high clouds with $\tau > 60$, which compensates for a lack of reduction of outgoing longwave radiation by thinner clouds at both middle and high levels of the troposphere. The progress is clearly identifiable but not quite as prominent as in the case of shortwave radiation with noticeable progress for the Community Atmosphere and Hadley Centre models but less so for the MIROC model and the multi-model means.

4. Scalar measures of the fidelity of model simulations

While the evidence just presented supports the notion that the simulation of clouds in climate models has been improving, it is helpful to provide scalar measures of the fidelity of model simulations that can quantitatively demonstrate progress. Here we present a few such quantities chosen to measure different aspects of cloud simulations and for which observational uncertainty is less than the differences between models and observations and among models themselves. These measures might be considered for a list of metrics for clouds in climate models [*Gleckler et al.*, 2008; *Pincus et al.*, 2008; *Williams and Webb*, 2009], although we do not develop this aspect here.

In the following, $c(ctp,\tau,X)$ is the amount of cloud in a given bin of the ISCCP histogram and is a function of cloud-top pressure *ctp*, optical depth τ, latitude, and generalized position X, including latitude, longitude, and month. Total cloud amount $C(\tau_{\min})$ is the sum of the cloud amounts of all bins with τ greater than the minimum optical thickness $\tau_{\min}$:

$$C(\tau_{\min},X) = \sum_{ctp} \sum_{\tau}^{\tau>\tau_{\min}} c(ctp,\tau,X) \tag{1}$$

We compute the normalized root-mean-square error Z_1 in the space-time distribution of total cloud amount, as:

$$Z_1(\tau_{\min}) = \sqrt{\int_X \left[C^{MOD}(\tau_{\min},X) - C^{OBS}(\tau_{\min},X)\right]^2} \Big/ \sigma_1 \,. \tag{2}$$

The integral in (2) denotes the area-weighted space-time average of squared differences between the model and ISCCP observations. The root-mean-square differences are normalized by the space-time standard deviation of the observed total cloud amount, given by:

$$\sigma_1 = \sqrt{\int_X \left[C^{OBS}(\tau_{\min}, X) - \bar{C}^{OBS}(\tau_{\min}) \right]^2} \,. \qquad (3)$$

As in Section 3.1, we set $\tau_{\min} = 1.3$.

Equation (1) uses the ISCCP simulator to ensure that model definitions of cloudiness are comparable with what is robustly observable but ignores the wealth of information provided by the joint histogram of *ctp* and τ. We evaluate the error Z_2 in this more finely-resolved distribution as the sum over a finite number of cloud-top pressure (N_{ctp}) and optical thickness (N_τ) bins of squared differences between the model and ISCCP observations:

$$Z_2 = \sqrt{\int_X \frac{1}{N_{ctp} \times N_\tau} \times \sum_{ctp} \sum_{\tau}^{\tau > \tau_{\min}} \left(c^{MOD}\left(ctp, \tau, X\right) - c^{OBS}\left(ctp, \tau, X\right) \right)^2} \Big/ \sigma_2 \,. \qquad (4)$$

This measure is sensitive to differences in each bin with $\tau > \tau_{\min}$, and would be applicable if the ISCCP simulator were capable of reproducing every aspect of the ISCCP observational processes. But comparisons with clouds retrieved from ground-based remote sensors and passed through the ISCCP simulator [Figures 2c and 3c of *Mace et al.*, 2011] suggest that the accuracy of ISCCP retrievals is about ±200 hPa for *ctp* and a factor of 3 for τ. We therefore compute Z_2 from a reduced-resolution histogram with bin boundaries in *ctp* of 440 hPa and 680 hPa and in τ of 3.6 and 23. (This is equivalent to the reduced-resolution joint histogram available in the monthly-averaged ISCCP data archives.) Considering the greater uncertainty of thin-cloud retrievals, we set $\tau_{\min} = 3.6$, and calculate differences only for the 6 bins with $\tau > \tau_{\min}$. Z_2 is normalized by σ_2, the accumulated space-time standard deviation of observed cloud amounts in the reduced bin set, making Z_2 the normalized root-mean-square error in the amount of low-intermediate, low-thick, medium-intermediate, medium-thick, high-intermediate, and high-thick clouds.

We compute the radiately-relevant error Z_3 in the distribution of clouds by using the radiative kernels to weight bin-by-bin errors by their radiative impact on top-of-atmosphere radiation fluxes:

$$Z_3^{SW,LW}(\tau_{\min}) = \sqrt{\int_X \frac{1}{N_{ctp} \times N_\tau} \times \sum_{ctp} \sum_{\tau}^{\tau > \tau_{\min}} \left[K^{SW,LW}(ctp,\tau,X) \times \left(c^{MOD}\left(ctp,\tau,X\right) - c^{OBS}\left(ctp,\tau,X\right) \right) \right]^2} \Big/ \sigma_3^{SW,LW} \tag{5}$$

Multiplication by radiative kernel is performed for each bin of the original ISCCP histogram before aggregation to the reduced bin set. This measure Z_3 has separate components for the shortwave and longwave spectrum, and is normalized by the accumulated space-time standard deviation of the radiative impacts of observed clouds from the reduced bin set.

Figure 10 shows Z_1, Z_2, Z_3^{LW}, and Z_3^{SW} for each model stratified into two rows according to the model ensemble. Arrows from earlier to later models indicate the change with time in the fidelity of model simulations; left-pointing arrows indicate smaller errors over time. The arrows connect the earliest and latest models from the modeling centers for which we track progress as well as the mean measure of each model ensemble. (In order to identify progress over time, the mean only includes the earliest CFMIP1 (latest CFMIP2) models from modeling centers that contribute more than one model to a given ensemble.)

For the total cloud amount measure Z_1, values range from 0.65 to 1.18 indicating that the standard deviation of biases in total cloud amount relative to ISCCP are generally comparable in size to the space-time of standard deviation of observed total cloud amount. To put this number into context, the Z_1 measure between the MODIS and ISCCP climatologies is 0.47. All model differences with ISCCP exceed this value, so it is likely that errors in the climatology of total cloud amount are robustly determined. Consistent with Figure 1, there is not a clear sign of improvement when considering the ensemble as a whole with the CFMIP1 ensemble mean value of Z_1 equal to 0.86 and the CFMIP2 ensemble mean value of Z_1 equal to 0.82. However, improvement is found for the Hadley Centre and Community Atmosphere models with a reduction of Z_1 from 1.12 for HadSM3 to 0.70 for HadGEM2A and a reduction of Z_1 from 0.94 for CCSM3.0 to 0.65 for CAM5, with little change in Z_1 for the Canadian and MIROC models or the ensemble mean.

For the cloud property measure Z_2, much more dramatic progress can be found. For three of the 4 models in which we can track progress (Hadley Centre, Community Atmosphere, and Canadian Centre models), errors relative to ISCCP has been reduced by 40-45% (relative), from 150-175% to 80-105% of the standard deviation of the ISCCP amounts of the 6 intermediate and thick cloud types. For the ensemble mean measure, more moderate progress can be found with 15-30% (relative) reduction in Z_2. Separate calculations reveal that the majority of the improvement in Z_2 comes from a better simulation of the amounts of optically intermediate ($3.6 < \tau < 23$) and thick ($\tau > 23$) clouds, than it does for improvements in the high, middle, and low amounts of clouds (with $\tau > 3.6$) (figures not shown). For the equivalent error measure calculated using only two bins for optically intermediate and thick clouds regardless of *ctp*, the value for the best model HadGEM2A is close to that calculated for differences between the observed ISCCP and MODIS distributions (0.70 vs. 0.59).

Radiatively-relevant cloud property measures Z_3^{SW} and Z_3^{LW} are shown in the bottom row of Figure 10. Similar to the cloud property measure Z_2, both measures show significant error reductions of 20-30% for the ensemble mean measure with larger 40-50% error reductions for individual models such as those of the Hadley Centre and Community Atmosphere. Again, the majority of this error reduction comes from improvement in the simulation of τ, indicating that models are better simulating the amount of shortwave radiation reflected and longwave radiation trapped by optically intermediate and thick clouds. Though it may appear that there is a redundancy among Z_2, Z_3^{SW} and Z_3^{LW}, only

Z_2 and Z_3^{SW} are highly correlated; all other possible pairings, including those with Z_1, have statistically insignificant inter-model correlations.

5. Why are simulations of clouds improving, and what impacts might this have?

The agreement between satellite observations and simulations by climate models of the climatological annual cycle of cloud amount, cloud-top pressure, and optical thickness has improved over the last decade. The improvement is most striking in the simulation of τ, where a bias of having too many optically thick clouds ($\tau > 23$) has been reduced by about 50% in the multi-model mean, with the best models having eliminated this bias. With a corresponding increase in the simulated amount of clouds with intermediate optical depth ($3.6 < \tau < 23$), this reduces the tendency for climate models to simulate approximately the right amount of shortwave radiation reflected by clouds but with the compensating errors of having too few clouds that are too bright.

Improvement in the amount or height distribution of clouds is not clear in the ensemble as a whole although progress can be found in individual models. For example, the simulations of total cloud amount in the Hadley Centre and Community Atmosphere models do show noticeable improvement (see Z_1 of Figure 10); in part, this improvement results from better simulations of the amount of clouds in the climatically important subtropical marine stratocumulus regions, where the amount of cloud is close to that observed in their most recent models. Some things show no improvement in the majority of climate models such as the underestimate of cloud over middle-latitudes, particularly over land, and an overestimate in the amount of optically thick cloud over tropical land.

Pinpointing the reasons for model improvement is difficult without testing of individual modifications from among the myriad of changes that modeling centers have implemented in the last decade, and it is likely that many factors have contributed. Even apart from parameterization changes, the incorporation of ISCCP simulator diagnostics in the routine evaluation of developmental model versions (as was done at the Hadley Centre for much of the last decade [*Martin et al.*, 2006]) can have a subtle but persistent influence on the choices made in the model-development process in such as way as to lead to improved simulation of clouds.

With regard to parameterizations, the improved boundary layer turbulence and shallow convection parameterizations in the Hadley Centre and Community Atmosphere models [*Lock et al.*, 2000; *Bretherton and Park*, 2009; *Park and Bretherton*, 2009] are almost certainly responsible for the improved simulations in marine stratocumulus clouds. However, in the case of the improved optical depth distribution, the causes of improvement are less clear but there are some clues from what has happened at the individual modeling centers whose progress we can track.

Beginning with the Canadian model, the reduction in the amount of optically thick cloud between its two versions is striking given the relatively few changes between model versions (J. Cole, personal communication). The likeliest cause is thought to be the introduction into CanAM4 of a new treatment of sub-grid scale variability in cloud optical properties known as the Monte Carlo Independent Column Approximation (McICA) [*Pincus et al.*, 2003]. The improvement treatment of cloud overlap and sub-grid scale heterogeneity in τ, with a retuning of the model, is apparently responsible for the reduction in optically thick cloud. In this example, an improved treatment of the radiative impact of clouds permitted better clouds properties to be simulated in a model that must match the observed radiation budget. A sensitivity study using McICA in the GFDL model [see Figure 4 of *Zhang et al.*, 2005] also shows a noticeable reduction in the amount of optically thick cloud.

In the Hadley Centre models, McICA is not used so other explanations must be sought. The largest reduction in optically thick cloud happened between HadSM3 and HadSM4, with a smaller but still sizeable reduction between HadSM4 and HadGSM1. Between HadSM3 and HadSM4, boundary layer vertical resolution was increased, the *Lock et al.* [2000] boundary layer turbulence parameterization was introduced, as was a sub-grid (in the vertical) treatment of cloud fraction. The possibility for clouds to occur in thinner layers admits the possibility of lower optical depths in stratiform clouds to be simulated (at fixed water content) (M. Webb, personal communication). The vertical resolution of climate models is known to be too coarse to simulate the many stratiform clouds that have geometrical cloud thicknesses smaller than that typical of model layers. Additionally, HadSM4 introduced an improved treatment of mixed-phase cloud microphysics [*Wilson and Ballard*, 1999] which also may be a factor in the reductions of optically thick cloud, particularly at middle-latitudes where a treatment of the Bergeron process may reduce the amount of super-cooled liquid in deep frontal clouds.

In the Community Atmosphere models, the vertical resolution in the boundary layer was increased and every physical parameterization, except that of deep convection, was changed between CAM4 and CAM5. Thus, all of the explanations above may be playing a role in their reduction of optically thick cloud [*Neale et al.*, 2011]. In particular, the introduction of improved cloud microphysics led to a substantial reduction in liquid water path over middle-latitudes that probably contributes to the reduction of optically thick clouds [*Gettelman et al.*, 2008].

Our evaluation is necessarily incomplete. For example, it is of interest to evaluate other cloud properties, such as liquid and ice water paths, or modes of variability, or how clouds co-vary with environmental parameters including 500 hPa vertical velocity and lower tropospheric stability. Because our analysis requires the use of an ISCCP simulator, our study is limited in the number of models that we can examine, although most major climate models have been included in this study. Evaluation of the limited and less consistently determined cloud information collected from a wider set of climate models is also of interest [*Jiang et al.*, 2012].

One may wonder if there is any connection between improved cloud simulations in climate models and the response to greenhouse gases in the climate changes these model simulate. Previous investigations have found no significant relationship between climate sensitivity and the fidelity of a model simulation in simulating present-day climate of clouds and precipitation [*Pincus et al.*, 2008]. We note that range of climate sensitivity in CMIP5 models is just as wide as it was in CMIP3 [*Andrews et al.*, 2012], again with the diversity in cloud feedbacks being a leading cause of inter-model spread. This suggests that there is no connection between the global mean cloud feedback and the fidelity with which a model simulates the clouds of the present-day climate. One implication of the reduction of cloud optical depths is that the magnitude of cloud feedbacks resulting from optical depth changes can be substantially larger if the current climate's cloud albedo is not saturated [*Stephens* 2010].

Acknowledgments. We acknowledge the World Climate Research Program's Working Group on Coupled Modeling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Tables 1 and 2 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The efforts of all authors from Lawrence Livermore National Laboratory were supported by the Regional and Global Climate and Earth System Modeling programs of the United States Department of Energy's Office of Science and were performed under the auspices of the United States Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. RP appreciates support from NASA under grant NNX11AF09G and from NSF under grant AGS 1138394. We thank Ben Sanderson for providing ISCCP simulator output from the CAM4 slab-ocean model, and Alejandro Bodas-Salcedo for providing additional ISCCP simulator output from the Hadley Center models. We thank Jason Cole, Mark Webb, and Shaocheng Xie for conversations.

References

Ackerman, T. P., and G. M. Stokes (2003), The Atmospheric Radiation Measurement Program, *Phys. Today*, *56*, 38–44.

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012), Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, *Geophys. Res. Lett.*, *39*, L09712, doi:10.1029/2012GL051607.

Bodas-Salcedo, A., et al. (2011), COSP: Satellite simulation software for model assessment, *Bull. Amer. Meteor. Soc.*, *92*, 1023–1043.

Bony, S., and J.-L. duFresne (2005), Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, *Geophys. Res. Lett.*, *32*, L20806, doi:10.1029/2005GL023851.

Bony, S., M. Webb, C. Bretherton, S. Klein, P. Siebesma, G. Tselioudis, and M. Zhang (2011), CFMIP: Towards a better evaluation and understanding of clouds and cloud feedbacks in CMIP5 models, *CLIVAR Exchanges, 56*, International CLIVAR Project

Office, Southampton, United Kingdom, 20-24.

Bretherton, C. S. and S. Park (2009), A new moist turbulence parameterization in the Community Atmosphere Model, *J. Clim.*, *22*, 3422-3448.

Collins, W. D. et al. (2006), The formulation and atmospheric simulation of the Community Atmosphere Model Version 3 (CAM3), *J. Clim.*, *19*, 2144-2161.

Collins, W. J. et al. (2008), *Evaluation of the HadGEM2 model*, Met Office Hadley Centre Technical Note no. HCTN 74, Met Office, FitzRoy Road, Exeter EX1 3PB, United Kingdom.

Curry, J. A., W. B. Rossow, D. Randall, and J. L. Schramm (1996), Overview of Arctic cloud and radiation characteristics, *J. Clim.*, *9*, 1731–1764.

Dai, A. (2006), Precipitation characteristics in eighteen coupled climate models, *J. Clim.*, *19*, 4605-4630.

Garay, M. J., S. P. de Szoeke, and C. M. Moroney (2008), Comparison of marine stratocumulus cloud top heights in the southeastern Pacific retrieved from satellites with coincident ship-based observations, *J. Geophys. Res.*, *113*, D18204, doi: 10.1029/2008JD009975.

Gates, W. L., et al. (1999), An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I), *Bull. Amer. Meteor. Soc.*, *80*, 29–55.

Gent, P. R. et al. (2011), The Community Climate System Model Version 4, *J. Clim.*, *24*, 4973-4991.

Gettelman, A., H. Morrison, and S. J. Ghan (2008), A new two-moment bulk stratiform cloud microphysics scheme in the Community Atmosphere Model (CAM3), Part II: Single-column and global results, *J. Clim.*, *21*, 3660-3679.

GEWEX Cloud System Science Team (1993), The GEWEX Cloud System Study (GCSS), *Bull. Amer. Meteor. Soc.*, *74*, 387–399.

GFDL GAMDT (2004), The new GFDL global atmosphere and land model AM2/LM2: Evaluation with prescribed SST simulations, *J. Clim.*, *17*, 4641-4673.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hourdin, F. et al. (2006), The LMDZ4 general circulation model: climate performance and sensitivity to parametrized physics with emphasis on tropical convection, *Clim. Dyn.*, *27*, 787-813.

IPCC (2007), *Climate Change 2007: The Physical Science Basis*, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller (eds.)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Jiang, J. et al. (2012), Evaluation of cloud and water vapor simulations in CMIP5 climate models using NASA "A-Train" satellite observations, *J. Geophys. Res.*, doi:10.1029/2011JD017237.

Kay, J., et al. (2012), Exposing global cloud biases in the Community Atmosphere Model (CAM) using satellite observations and their corresponding instrument simulators, *J. Clim.*, in press.

Klein, S. A., and C. Jakob (1999), Validation and sensitivities of frontal clouds simulated by the ECMWF model, *Mon. Weather Rev.*, *127*, 2514–2531.

Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith (2000), A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests, *Mon. Wea. Rev.*, *128*, 3187–3199.

Ma, C.-C., C. R. Mechoso, A. W. Robertson, and A. Arakawa (1996), Peruvian stratus clouds and the tropical pacific circulation: A coupled ocean-atmosphere GCM study, *J.*

Clim., *9*, 1635–1645.

Mace, G. G., Q. Zhang, M. Vaughan, R. Marchand, G. Stephens, C. Trepte, and D. Winker (2009), A description of hydrometeor layer occurrence statistics derived from the first year of merged Cloudsat and CALIPSO data, *J. Geophys. Res.*, *114*, D00A26, doi:10.1029/2007JD009755.

Mace, G. G., S. Houser, S. Benson, S. A. Klein and Q. Min (2011), Critical evaluation of the ISCCP simulator using ground-based remote sensing data, *J. Clim.*, *24*, 1598–1612.

Marchand, R., T. Ackerman, M. Smyth, and W. B. Rossow (2010), A review of cloud top height and optical depth histograms from MISR, ISCCP, and MODIS, *J. Geophys. Res.*, *115*, D16206, doi:10.1029/2009JD013422.

Martin, G. M., et al. (2006), The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model (HadGEM1). Part I: Model description and global climatology, *J. Clim.*, *19*, 1274-1301.

McAvaney, B. J., and H. Le Treut (2003), The cloud feedback intercomparison project: (CFMIP). *CLIVAR Exchanges, 26*, International CLIVAR Project Office, Southampton, United Kingdom, 1-4.

Meehl, G., C. Covey, T. L. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, and R. J.

Stouffer and K. E. Taylor (2007), The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Amer. Meteor. Soc.*, *88*, 13830-1394.

Neale, R. B. et al. (2011a), *Description of the NCAR Community Atmosphere Model (CAM5)*, Technical Report NCAR/TN-486+STR, National Center for Atmospheric Research, Boulder, Colorado, U. S. A., 268 pp.

Ogura, T. et al. (2008), Towards understanding cloud response in atmospheric GCMs: The use of tendency diagnostics, *J. Met. Soc. Japan*, *86*, 69-79.

Park, S. and C. S. Bretherton (2009), The University of Washington shallow convection and moist turbulence schemes and their impact on climate simulations with the Community Atmosphere Model, *J. Clim.*, *22*, 3449-3469.

Pincus, R., H. W. Barker, and J. Morcrette (2003), A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous clouds, *J. Geophys. Res.*, *108*(D13), 4376, doi:10.1029/2002JD003322.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Gleckler (2008), Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, *J. Geophys. Res.*, *113*, D14209, doi:10.1029/2007JD009334.

Pincus, R., S. Platnick, S. A. Ackerman, R. S. Hemler, R. J. P. Hoffmann (2012), Reconciling simulated and observed views of clouds: MODIS, ISCCP, and the limits of instrument simulators, *J. Clim.*, *25*, 4699-4720.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton (2000), The impact of new physical parametrizations in the Hadley Centre climate model – HadAM3, *Clim. Dyn.*, *16*, 123-146.

Raddatz, T. J. et al. (2007), Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty first century? *Clim. Dyn.*, *29*, 565-574, doi:10.1007/s00382-007-0247-8.

Rossow, W. B. and R. A. Schiffer (1991), International Satellite Cloud Climatology Project (ISCCP) cloud data products, *Bull. Amer. Meteor. Soc.*, *72*, 2–20.

Rossow, W. B. and R. A. Schiffer (1999), Advances in understanding clouds from ISCCP, *Bull. Amer. Meteor. Soc.*, *80*, 2261–2288.

Slingo, A., and J.-M. Slingo (1988), The response of a general circulation model to cloud longwave radiative forcing. I. Introduction and initial experiments, *Quart. J. Roy. Met. Soc.*, *114*, 1027-1062.

Stephens, G. L., et al. (2002), The Cloudsat mission and the A-train, *Bull. Amer. Meteor.*

Soc., *83*, 1771–1790.

Stubenrauch, C., S. Kinne, and the GEWEX Cloud Assessment Team (2009), Assessment of global cloud climatologies, *GEWEX Newsletter*, *19*, International GEWEX Project Office, Silver Spring, Maryland, Unites States of America, 6-7.

Stephens, G. (2010), *Is there a missing low-cloud feedback in current climate models?* GEWEX Newsletter, 20, International GEWEX Project Office, Silver Spring, Maryland, United States of America, 5-7.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experimental design, *Bull. Amer. Meteor. Soc.*, *93*, 485-498.

Thorsen, T. J., Q. Fu, and J. Comstock (2011), Comparison of the CALIPSO satellite and ground-based observations of cirrus clouds at the ARM TWP sites, *J. Geophys. Res.*, *116*, D21203, doi:10.1029/2011JD015970.

Voldoire, et al. (2012), The CNRM-CM5.1 global climate model: description and basic evaluation, *Clim. Dyn.*, doi:10.1007/s00382-011-1259-y.

Webb, M., C. Senior, S. Bony, and J. J. Morcrette (2001), Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models, *Clim. Dyn.*, *17*, 905–922.

Williams, K. D. and M. J. Webb (2009), A quantitative performance assessment of cloud regimes in climate models. *Clim. Dyn.*, *33*, 141-157.

Wilson, D. R. and S. P. Ballard (1999), A microphysically based precipitation scheme for the U. K. Meteorological Office Unified Model, *Q. J. Roy. Met. Soc.*, *125*, 1607-1636.

Winker, D., et al. (2009), Overview of the CALIPSO mission and CALIOP data processing algorithms, *J. Atmos. Oceanic Technol.*, *26*, 2310-2323.

Yang, G.-Y. and J. Slingo, (2001), The diurnal cycle in the tropics, *Mon. Wea. Rev.*, *129*, 784-801.

Yukimoto, S. et al. (2011a), *Meteorological Research Institute – Earth System Model Version 1 (MRI-ESM1): Model Description*, Technical Report #64, Meteorological Research Institute, Tsukuba-city, Ibaraki 305-0052, Japan, 96 pp.

Zelinka, M. D., S. A. Klein and D. L. Hartmann (2012), Computing and partitioning cloud feedbacks using cloud property histograms. Part I: Cloud radiative kernels, *J. Clim.*, *25*, 3715–3735.

Zhang, M. H., et al. (2005), Comparing clouds and their seasonal variations in 10 atmospheric general circulation models with satellite measurements, *J. Geophys. Res.*, *110*, D15S02, doi:10.1029/2004JD005021.

Tables

Table 1. CFMIP 1 slab ocean models used in this study.

Model Name	Modeling Center	Reference	Number of Years in Run	Symbol
AGCM4.0	Canadian Centre for Climate Modeling and Analysis	http://www.ec.gc.ca/ccmac-cccma/	20	c4
CCSM3.0	National Center for Atmospheric Research	*Collins et al.* [2004]	20	n3
GFDL MLM 2.1	NOAA / Geophysical Fluid Dynamics Laboratory	*GFDL GAMDT* [2004]	20	g
HadGSM1	Met Office Hadley Centre	*Martin et al.* [2006]	20	h1
HadSM3	Met Office Hadley Centre	*Pope et al.* [2000]	20	h3
HadSM4	Met Office Hadley Centre	*Webb et al.* [2001]	20	h4
IPSL CM4	Institut Pierre Simon Laplace	*Hourdin et al.* [2006]	20	i
MIROC (hisens)	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change	*Ogura et al.* [2008]	5	m3
MIROC (losens)	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change	*Ogura et al.* [2008]	5	m4

Table 2. CFMIP 2 AMIP models used in this study.

Model Name	**Modeling Center**	**Reference**	**Number of Years in Run**	**Symbol**
CAM4	Community Earth System Model Contributors (NSF-DOE-NCAR)	*Gent et al.* [2004]	10	N4
CAM5	Community Earth System Model Contributors (NSF-DOE-NCAR)	*Neale et al.* [2011]	10	N5
CanAM4	Canadian Centre for Climate Modeling and Analysis	http://www.ec.gc.ca/ccmac-cccma/	60	C4
CNRM-CM5	Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	*Voldoire et al.* [2012]	30	Q
HadGEM2A	Hadley Centre for Climate Prediction and Research/Met Office	*Collins et al.* [2008]	30	H2
MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	*Watanabe et al.* [2010]	30	M5
MPI-ESM-LR	Max Planck Institute for Meteorology	*Raddatz et al.* [2007]	30	P
MRI-CGCM3	Meteorological Research Institute	*Yukimoto et al.* [2011]	30	R

Figures

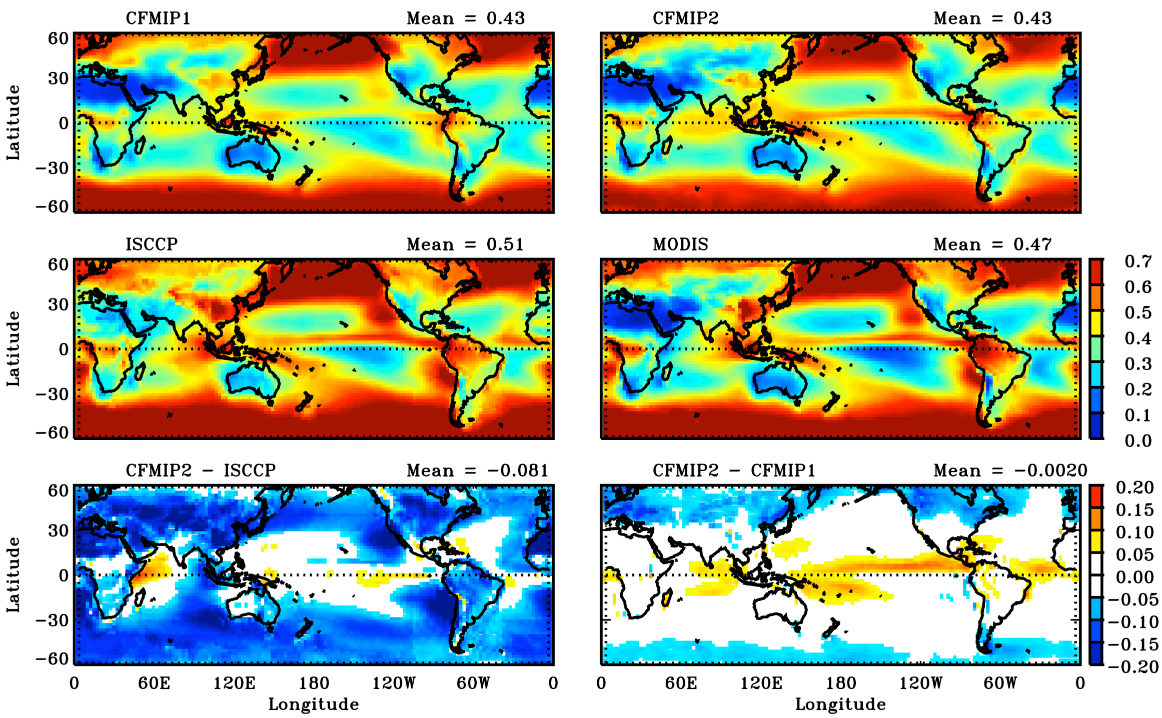

Figure 1. Total cloud amount ($\tau > 1.3$) from CFMIP1 and CFMIP2 multi-model means, ISCCP and MODIS observations, and the difference of CFMIP2 multi-model mean to the ISCCP and CFMIP1 multi-model mean.

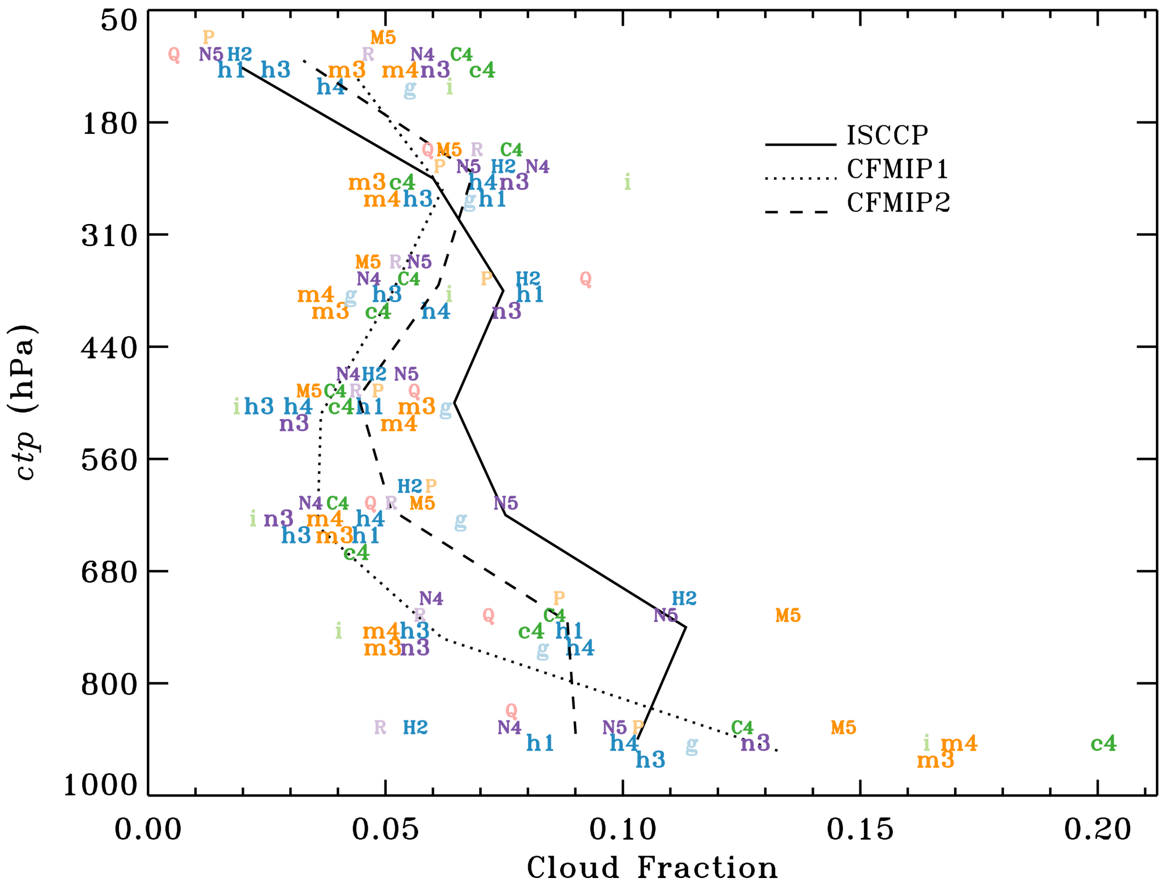

Figure 2. Fractional area in the domain 60ºS - 60ºN covered by clouds as a function of cloud-top pressure from models and ISCCP observations. CFMIP1 (2) ensemble means are plotted with a dotted (dashed) line. The area is computed only for clouds with τ > 1.3. The symbol key for models is provided in Tables 1 and 2.

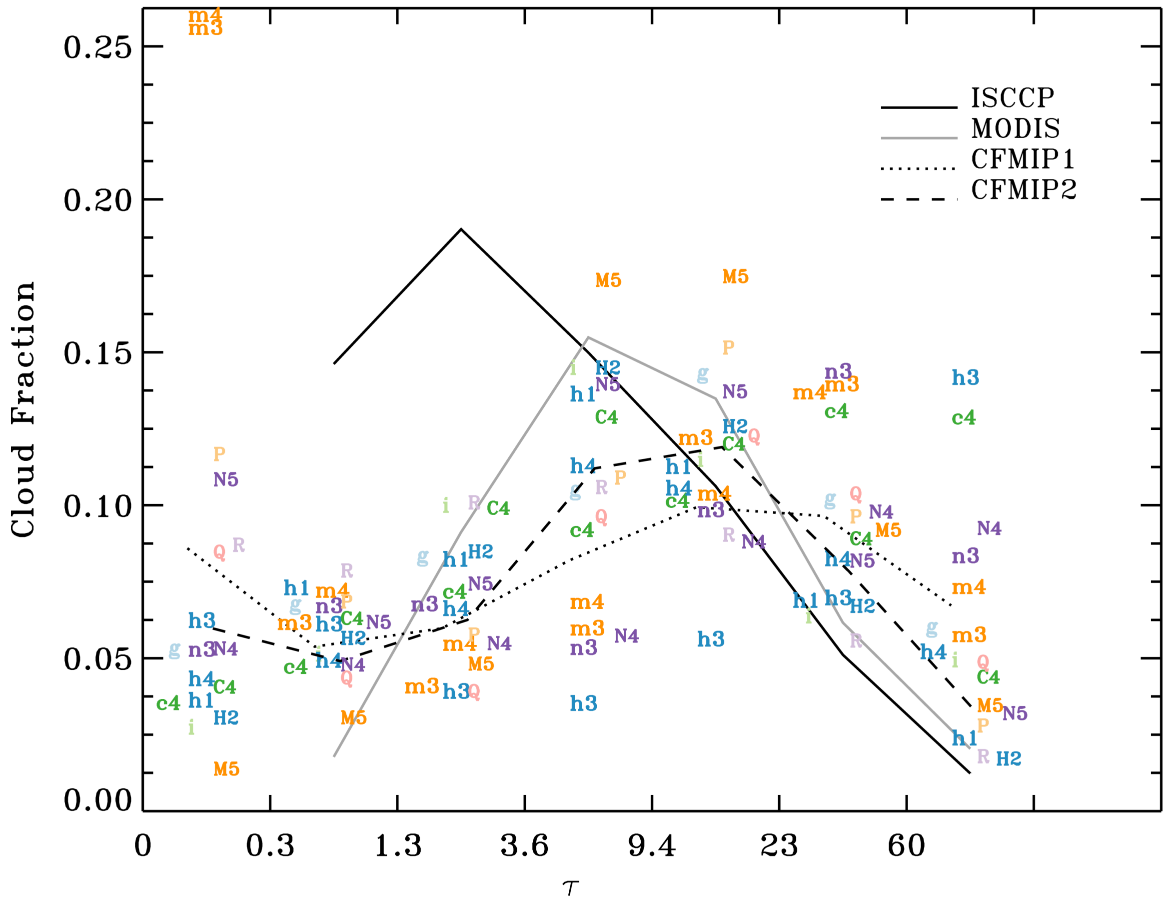

Figure 3. Fractional area in the domain 60ºS - 60ºN covered by clouds as a function of optical thickness from models and ISCCP and MODIS observations.

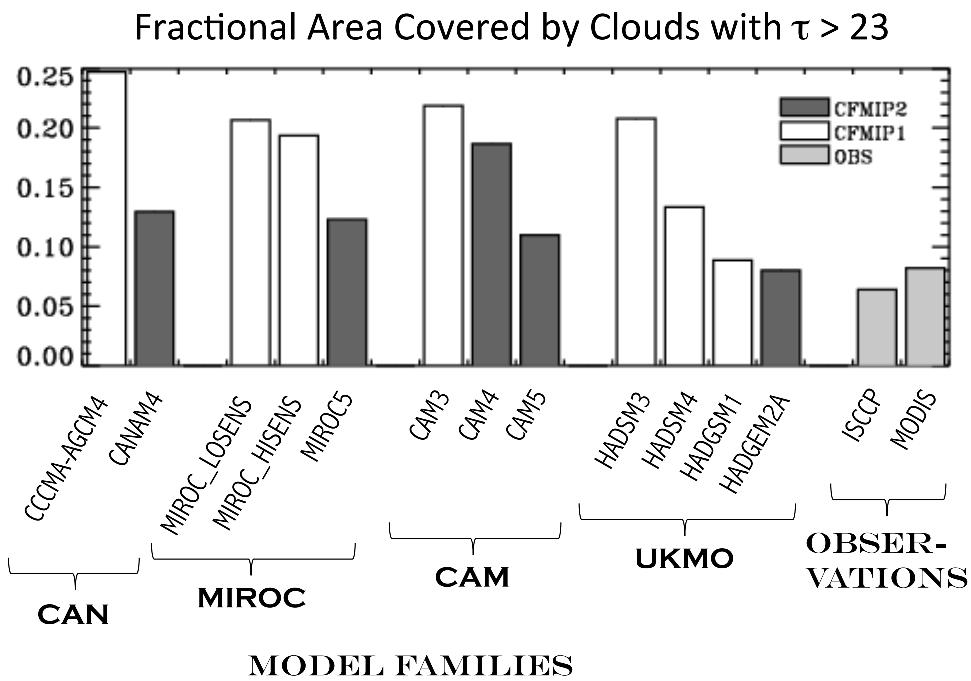

Figure 4. Fractional area in the domain 60ºS - 60ºN covered by clouds with $\tau > 23$ for selected model families and observations. Models are plotted so as to illustrate progress in reducing the overestimate of optically thick cloud over time by ordering models from earliest to latest (left to right) within families.

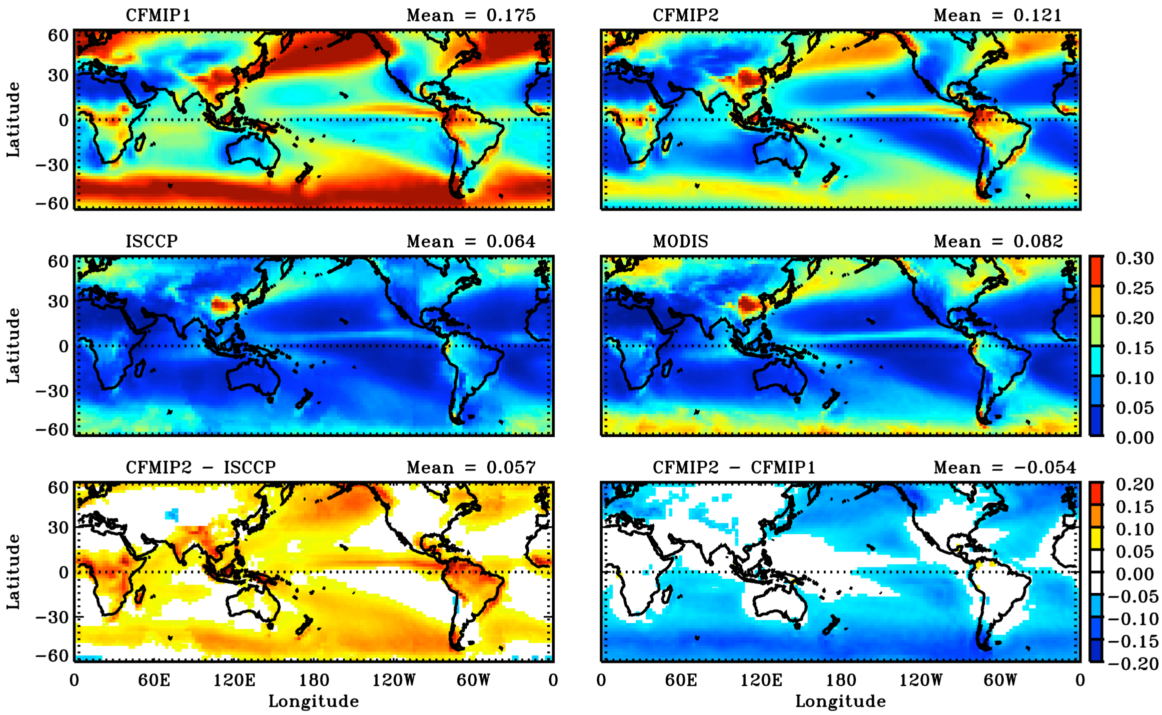

Figure 5. Fractional area covered by optically thick clouds ($\tau > 23$) from CFMIP1 and CFMIP2 multi-model means, ISCCP and MODIS observations, and the difference of the CFMIP2 multi-model mean to ISCCP and the CFMIP1 multi-model mean.

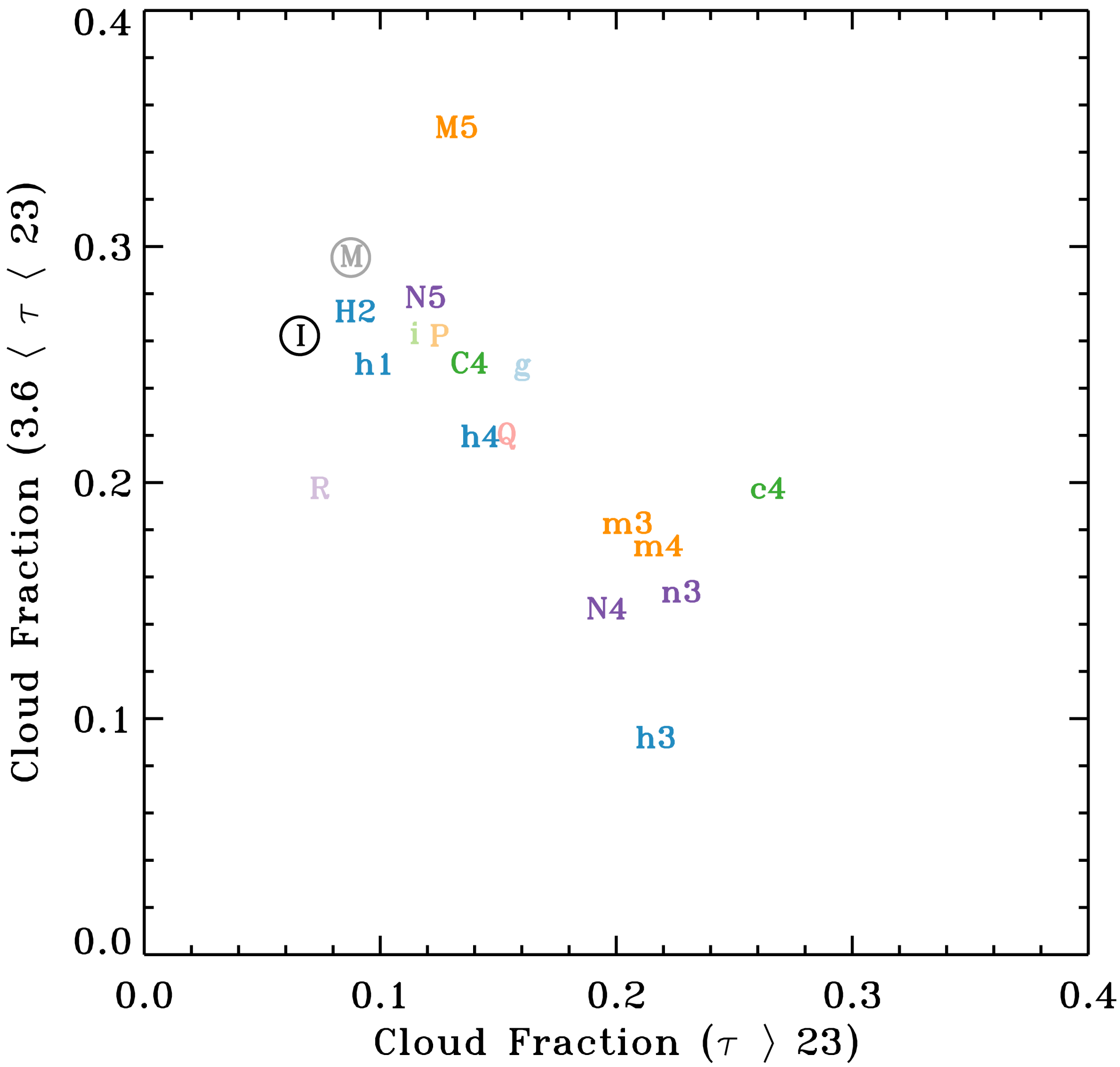

Figure 6. Scatterplot of the fractional area in the domain 60ºS - 60ºN covered by clouds with $\tau > 23$ and clouds with $3.6 < \tau < 23$. Observations from MODIS and ISCCP are represented by “M” and “I”, respectively. The symbol key for models is provided in Tables 1 and 2.

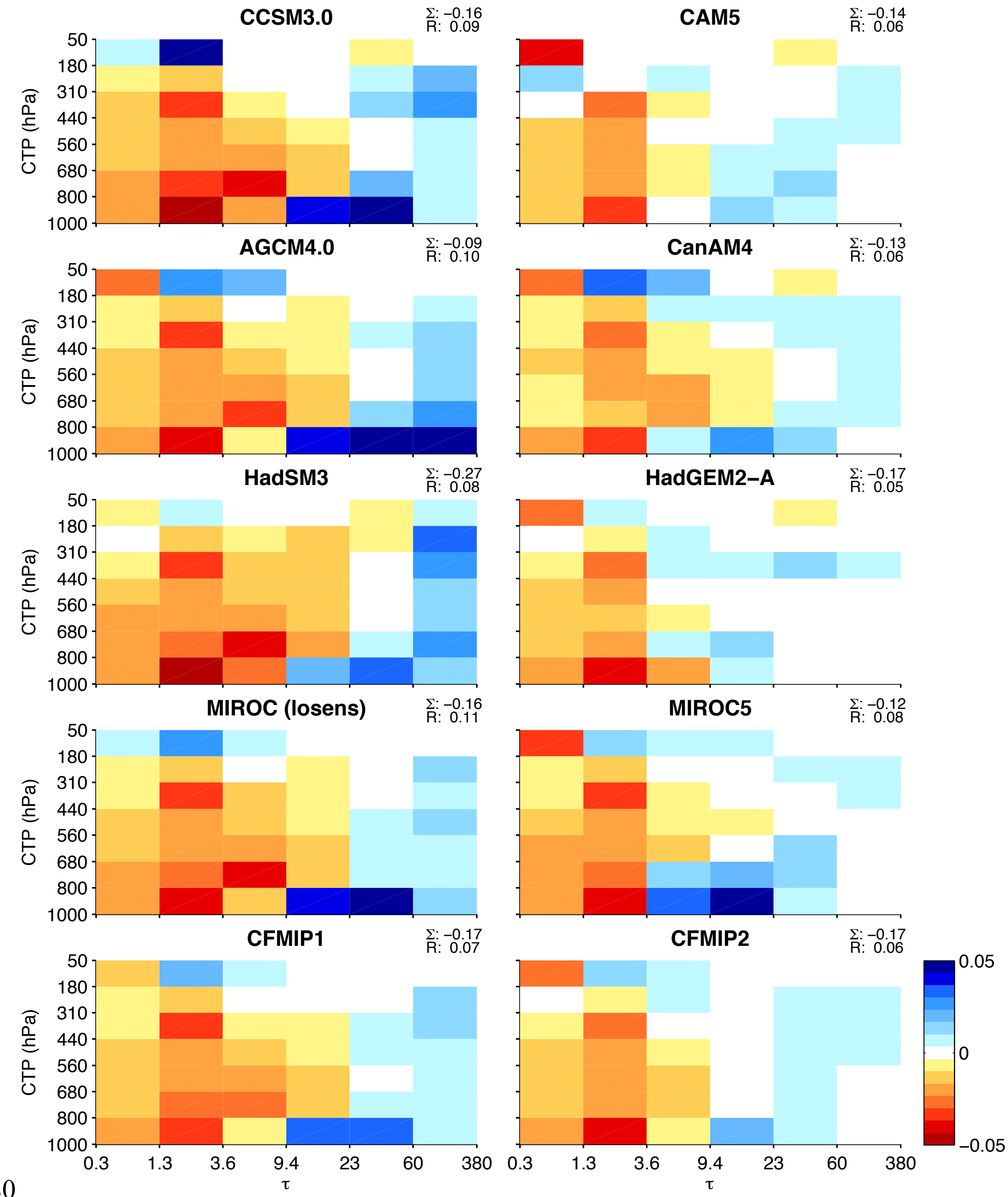

Figure 7. Area-averaged biases in the domain 60ºS - 60ºN with respect to ISCCP observations of fractional area covered by clouds in bins of cloud-top pressure and optical depth. Results are plotted for the 4 model families in which we track progress and the ensemble mean. Models are ordered with the oldest models on the left and the newest models on the right. The sum of the histogram and the range (maximum minus minimum value in the histogram) are shown in the title of each panel. Positive values indicate model overestimates relative to observations. The fact that the recent models have fewer bins with color as well as reduced intensity in the bins with color indicates improvements with time.

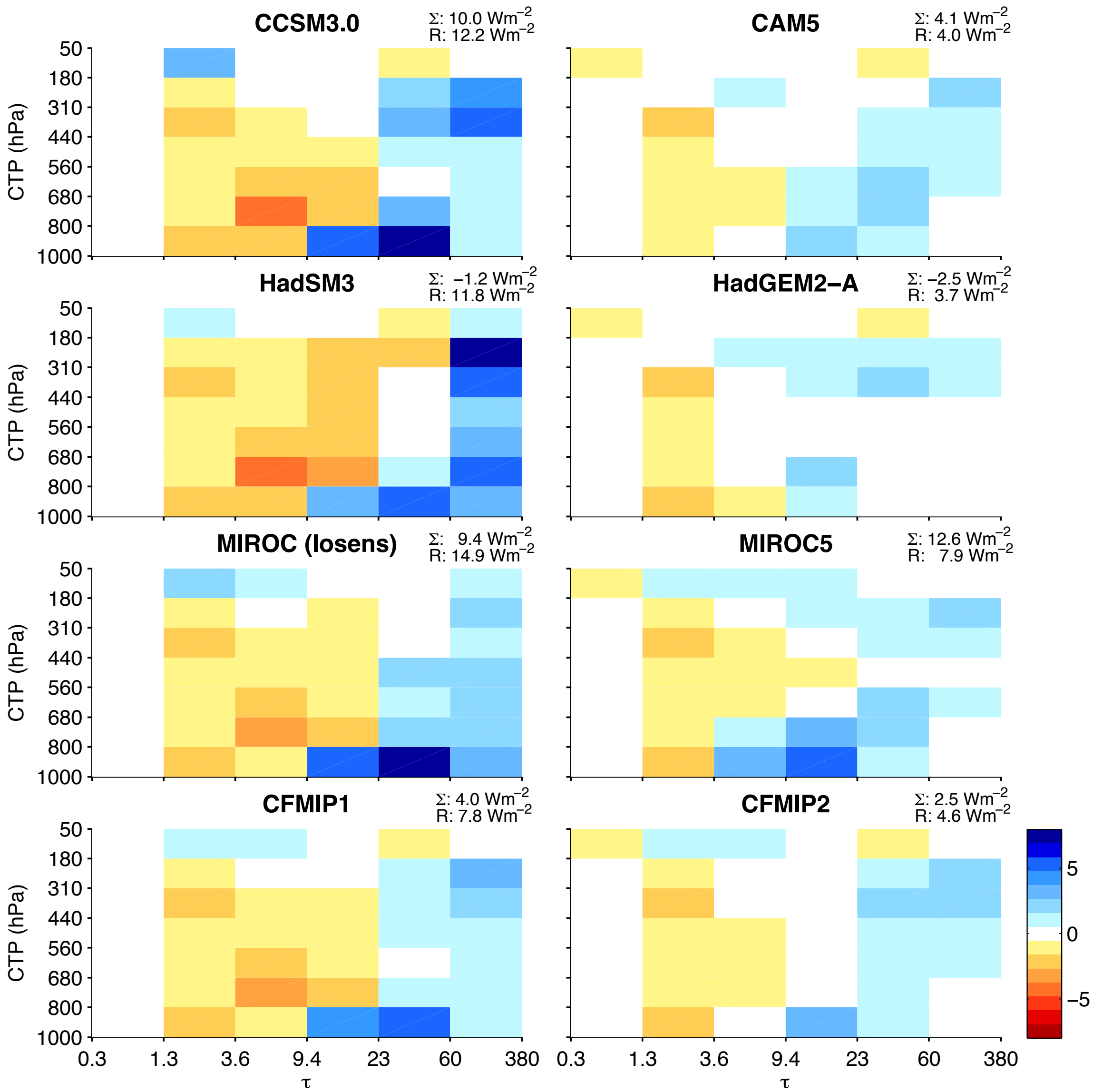

Figure 8. As in Figure 7, but for the contributions to shortwave radiation reflected to space by clouds stratified into bins of cloud-top pressure and optical depth. Positive values indicate a bias towards too much reflected radiation due to a positive bias in cloud amount.

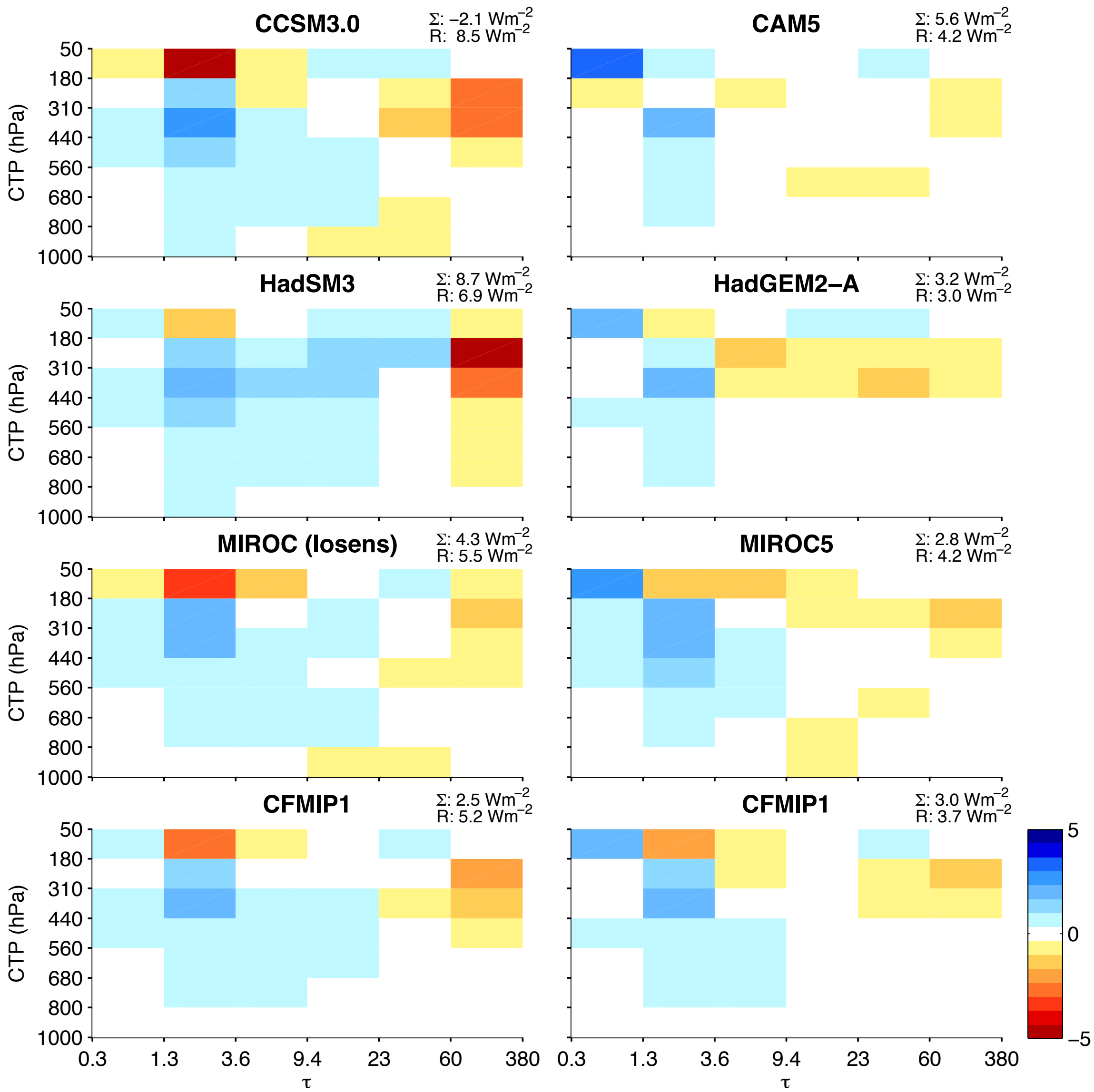

Figure 9. As in Figure 7, but for the contributions to reductions of outgoing longwave radiation (relative to clear-sky) by clouds stratified into bins of cloud-top pressure and optical depth. Positive values indicate a bias towards too much longwave radiation emitted to space due to a negative bias in cloud amount.

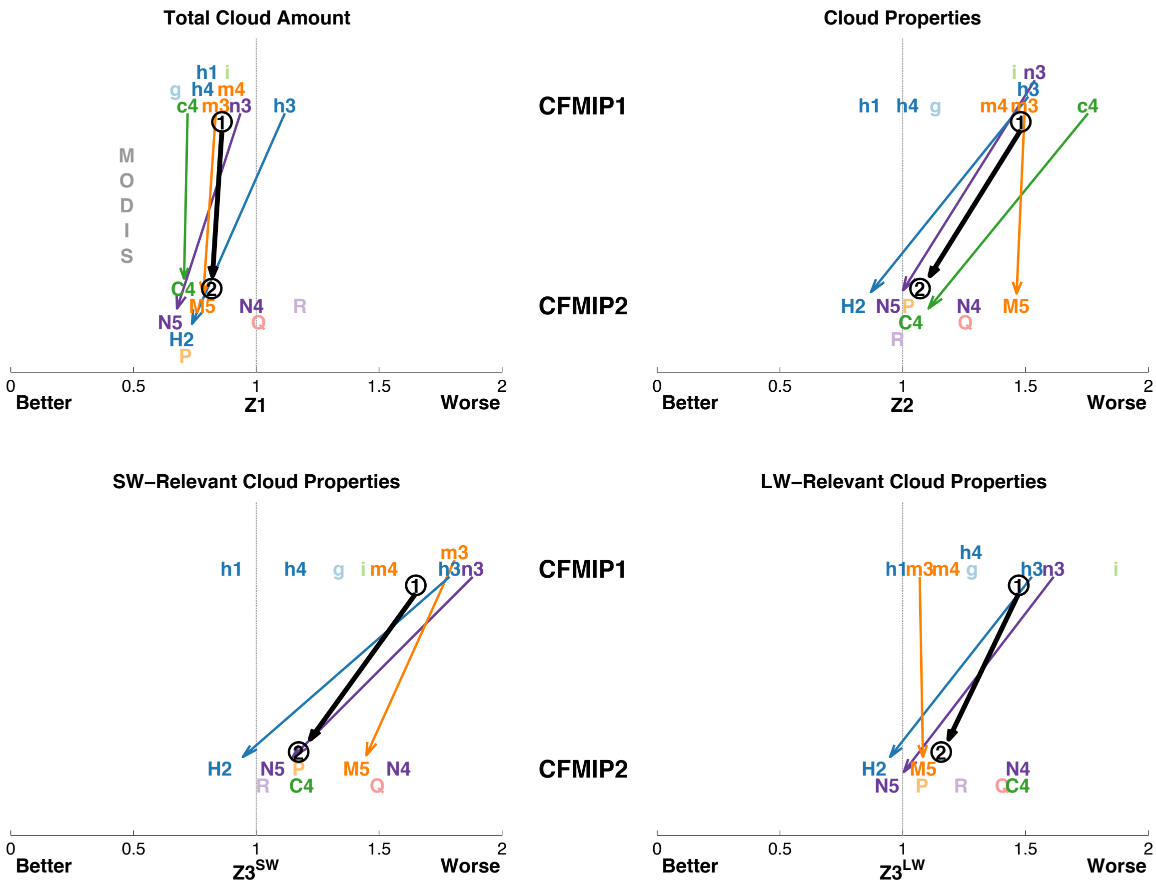

Figure 10. Scalar measures of fidelity of CFMIP model simulations in reproducing the space-time distribution of several cloud measures, with greater fidelity indicated by lower Z values. Z_1 measures fidelity in simulating total cloud amount, whereas Z_2 measures fidelity in simulating cloud-top pressure and optical depth in different categories of optically intermediate and thick clouds at high, middle, and low-levels of the atmosphere. Z_3 measures the impacts on top-of-atmosphere shortwave (lower left) and longwave (lower right) radiation in the same categories measured by Z_2. Models are stratified vertically into the two ensembles and are plotted according to the symbol key in Tables 1 and 2. For the modeling centers in which we can track progress, the arrow connects the oldest model in the family (arrow base) to the most recent model (arrow tip). The thick black arrow connects the average measure of CFMIP1 models (arrow base) to that of CFMIP2 models (arrow tip). Arrows pointing to the left indicate improvements with time.